MLDS CENTER
Maryland Longitudinal Data System
Better Data • Informed Choices • Improved Results

Balancing Data Privacy and Utility: Benefits and Challenges in Developing a Synthetic Version of the Maryland SLDS

NCES STATS-DC
July 26th, 2019

Outline

- Overview of MLDSC
- Overview of Synthetic Data Project
- Evaluation of Research Utility
- Evaluation of Disclosure Risk

The MLDSC

- Independent state agency
- Receives, matches and merges education and workforce data from 3 partner state agencies: MSDE (grades K-12), MHEC (postsecondary), & DLLR (wages)
- Mission: Produce research reports and dashboards to inform state policy, programming, and the public
- Data are confidential, sensitive, and PII:
  - Data confidentiality protected by federal and state laws
  - Access to data granted to MLDSC staff only

The Synthetic Data Project

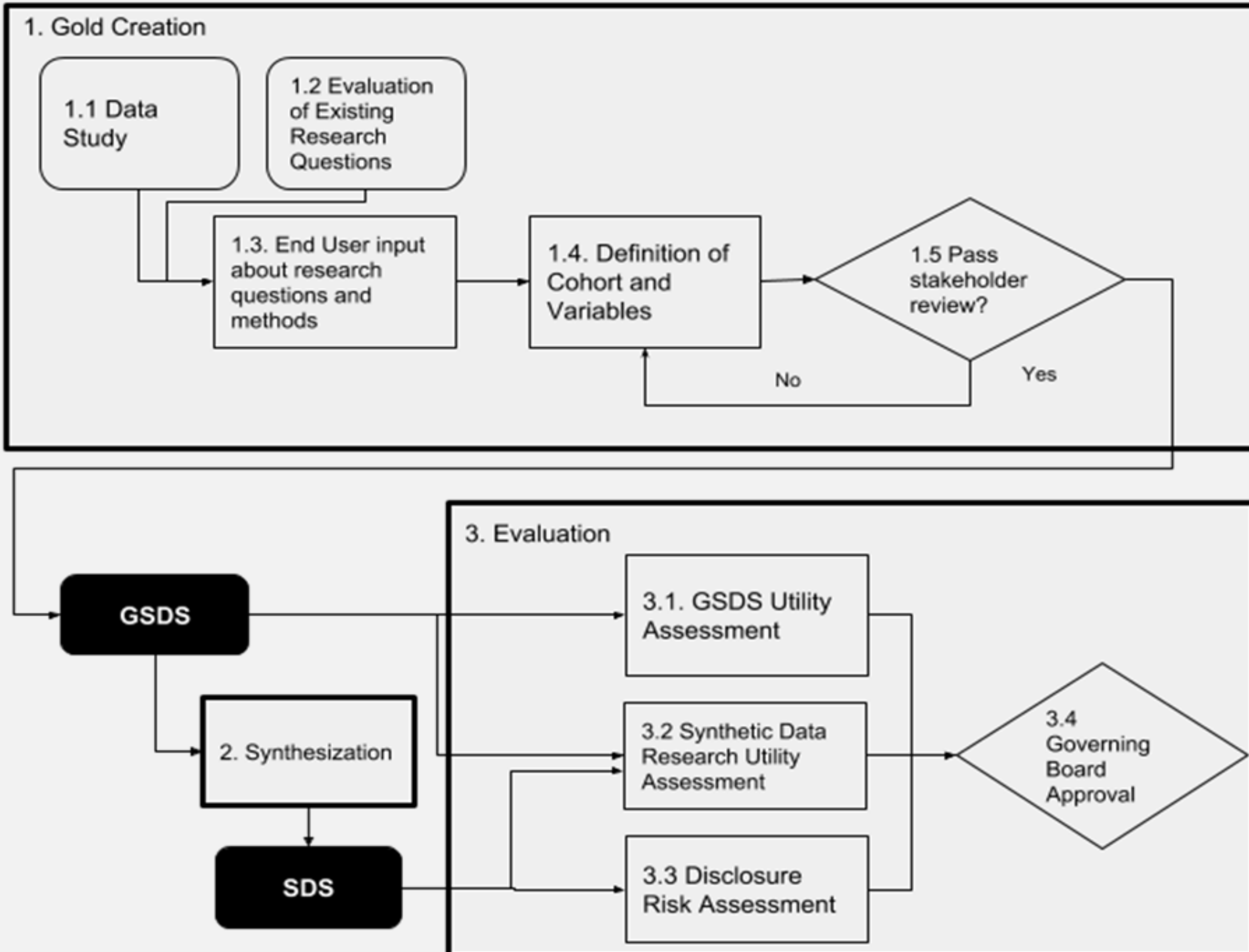
- 2015 SLDS grant from the U.S. Department of Education, Institute of Education Sciences ($2.7M) to create synthetic data version of the MLDS data. Aims:
    1. Create strategy to balance data confidentiality with need to make data available, and
    2. Expand access to the data to leverage research value.
- Synthetic data are generated based on models to mimic the relational patterns among variables
- Statistical analyses yield findings substantially similar to the real data, and
- simultaneously reduce the risk of privacy breach.

The Synthetic Data Project

1. Create gold standard datasets (GSDS)
   1. Study the data and potential uses
   2. Define GSDS
2. Synthesize GSDS
3. **Evaluate research utility and disclosure risk**

Next Steps:

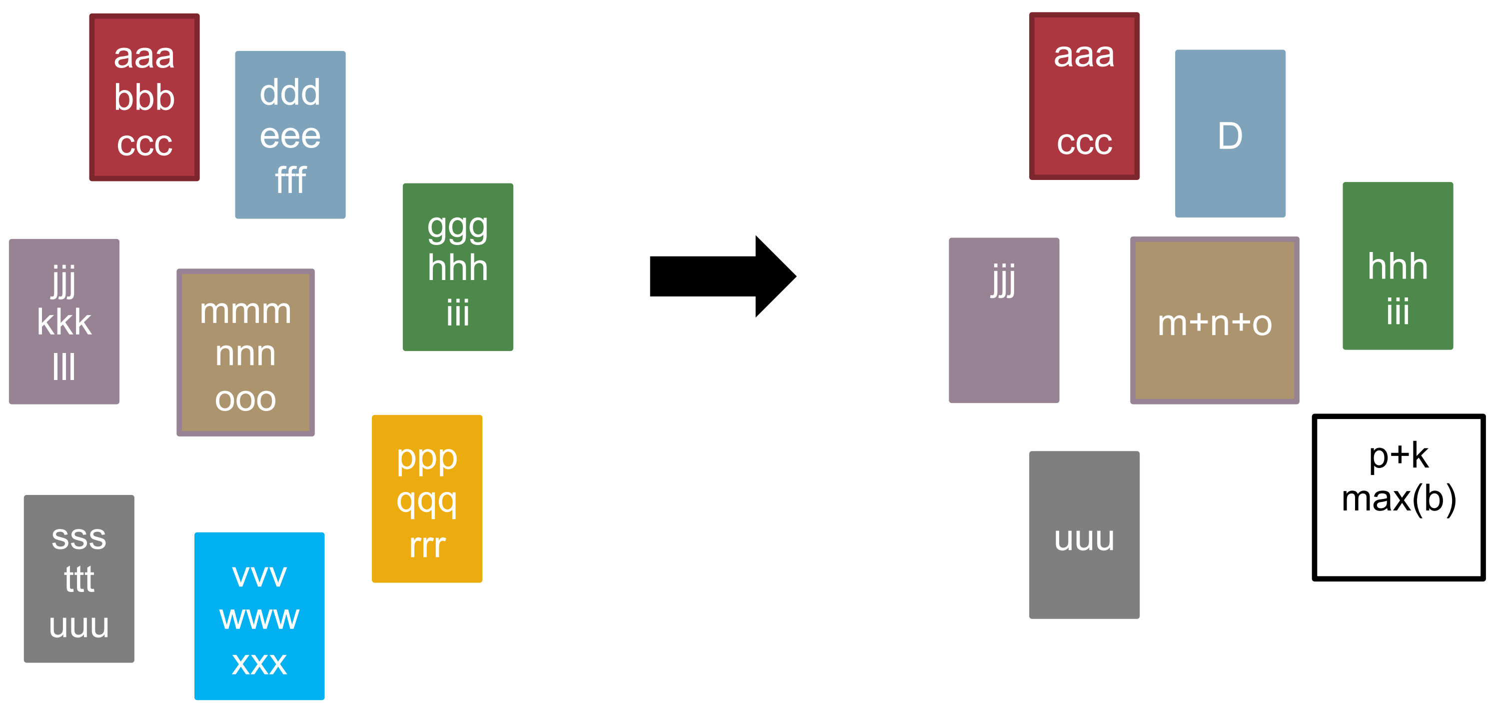
1. Governing Board Approval
   1. Beta Testing – RU & DR
   2. Release synthetic data
2. Report on the project to inform other state longitudinal data systems

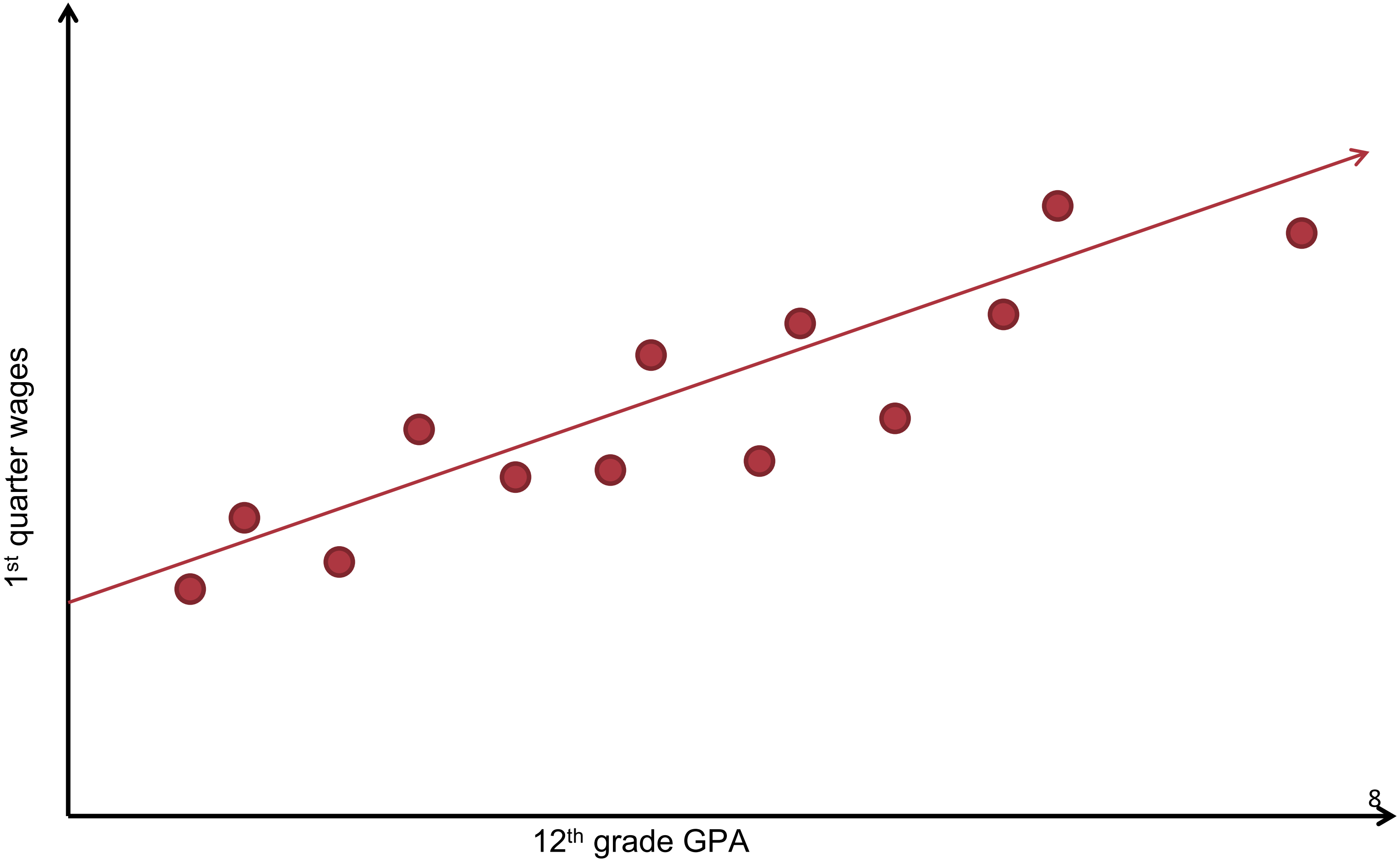

1. Gold Creation
1.1 Data Study
1.2 Evaluation of Existing Research Questions
1.3. End User input about research questions and methods
1.4. Definition of Cohort and Variables
1.5 Pass stakeholder review?
No
Yes
GSDS
2. Synthesization
SDS
3. Evaluation
3.1. GSDS Utility Assessment
3.2 Synthetic Data Research Utility Assessment
3.3 Disclosure Risk Assessment
3.4 Governing Board Approval

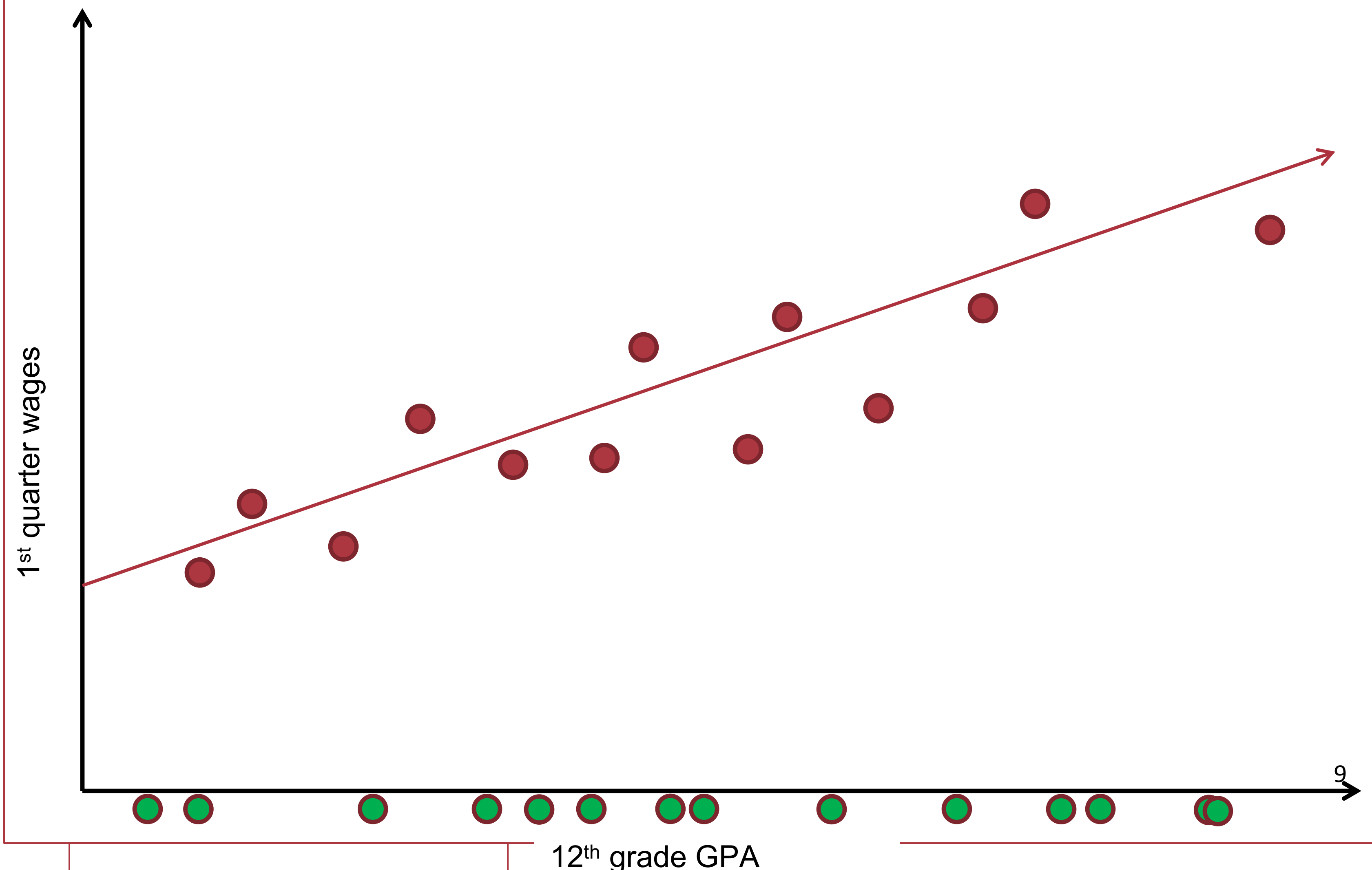
Creating the GSDS

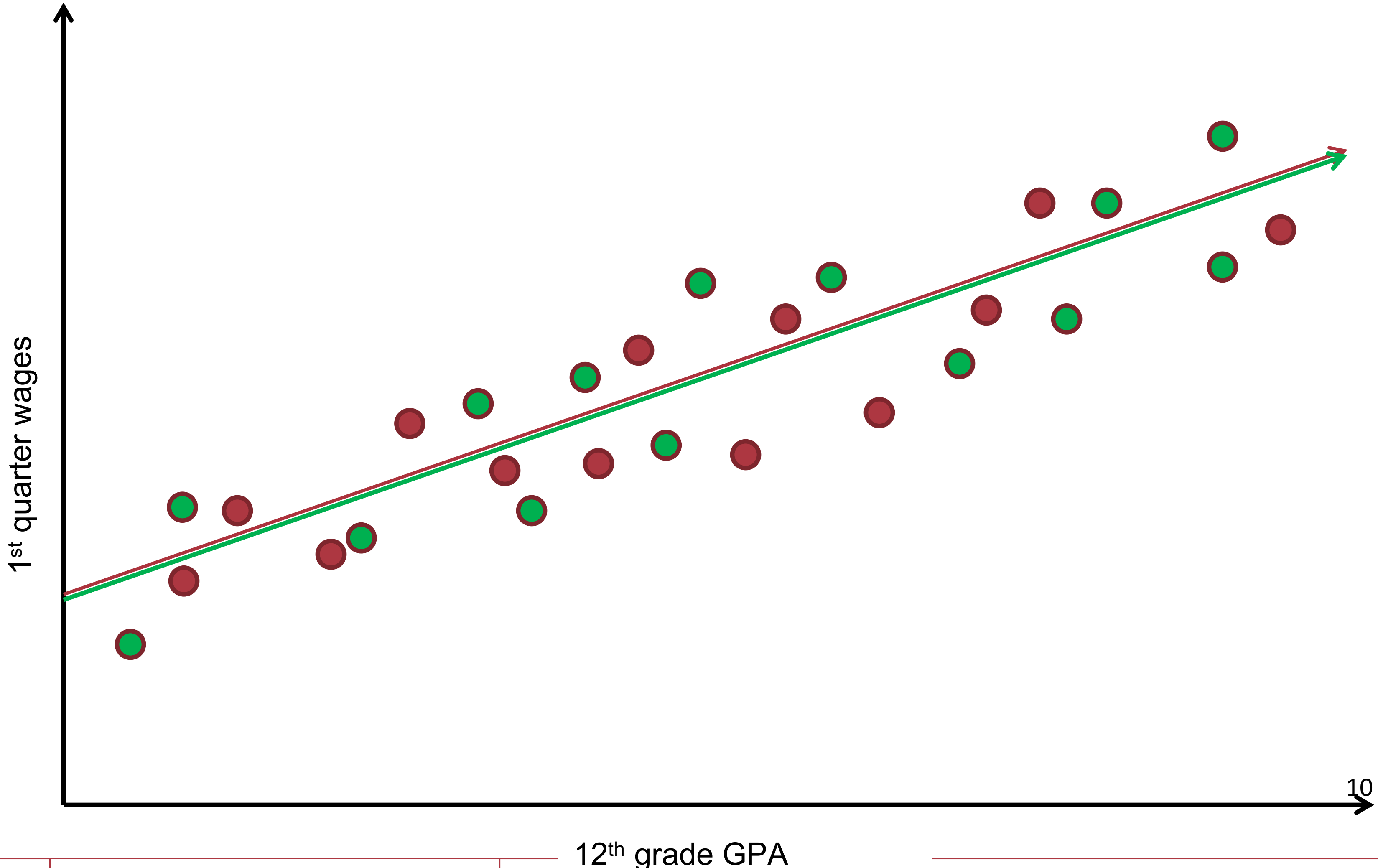
<u>Operational Data Store (ODS) (v=460)</u>

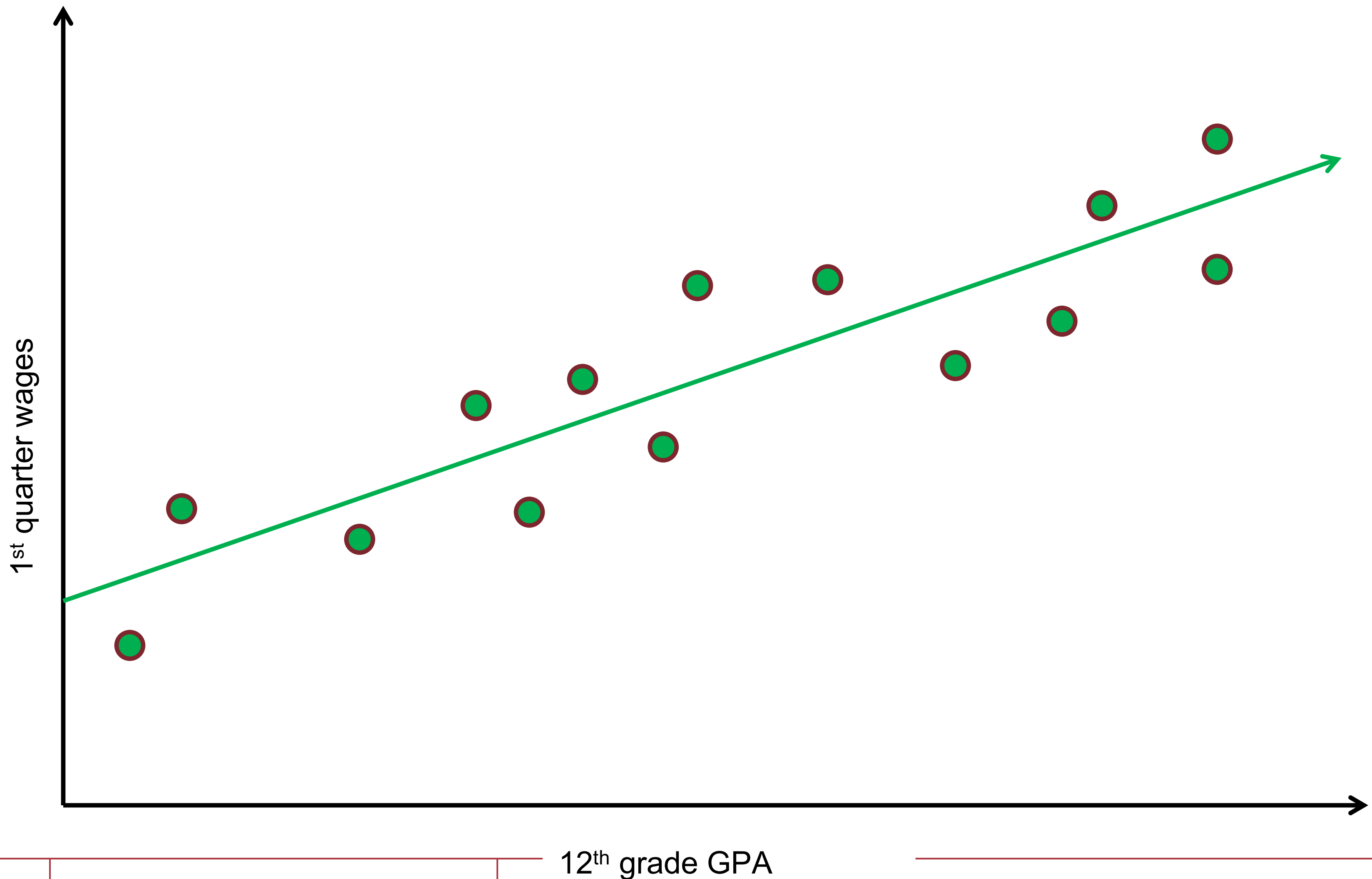
<u>Gold Standard Data Set (GSDS) (v=65, 50, 55)</u>
(But there are many rows of data per person!)




1st quarter wages
12th grade GPA

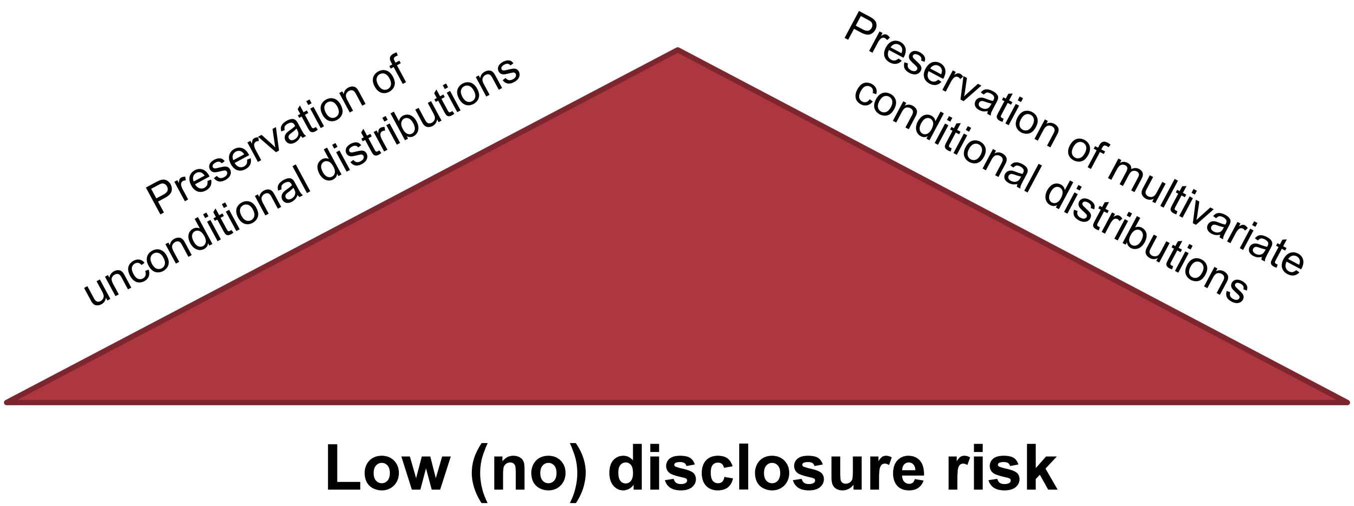

1st quarter wages
12th grade GPA


1st quarter wages
12th grade GPA


1st quarter wages
12th grade GPA

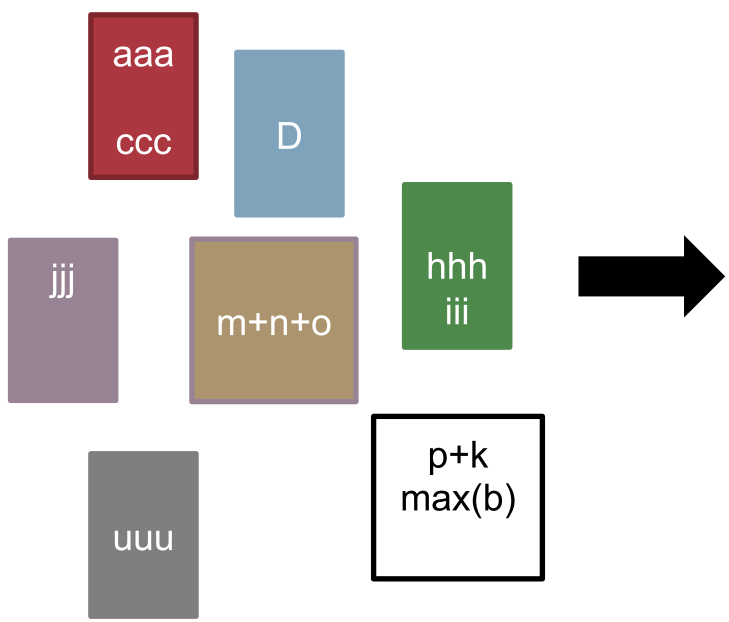
Synthesization (Step 2)

- We need to satisfy a triangular trade-off:



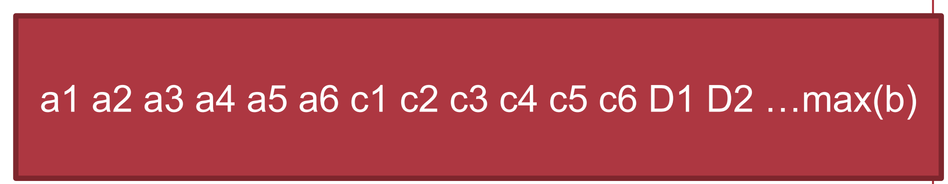
Synthesization

Gold Standard Data Set (GSDS) (v=65, 50, 55)



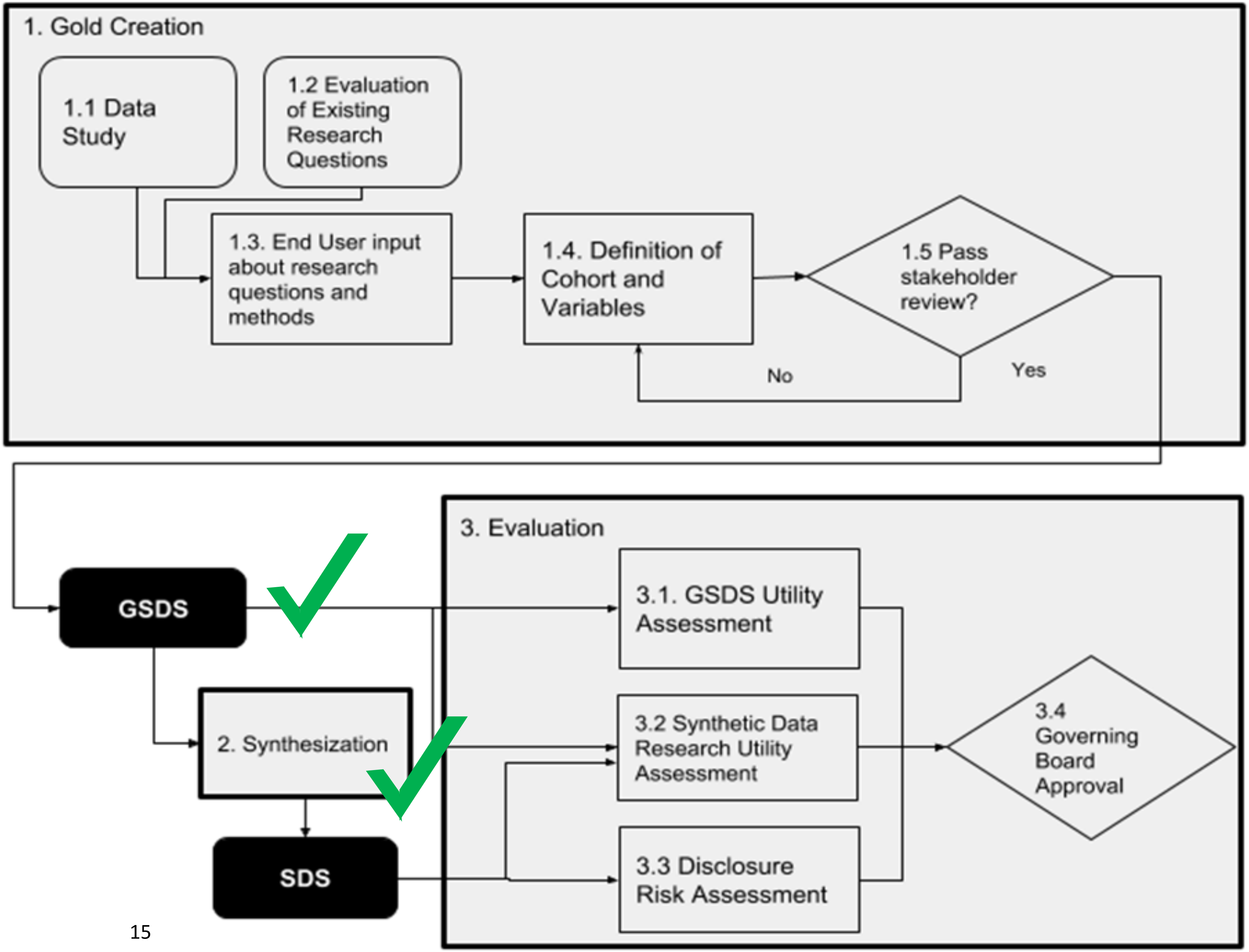
Transformed (v=4000, 4700, 5900)

For synthesis, we need one wide record per individual



Synthesization

- Given the number of variables, potential interactions and non-linearities, and after initial testing and evaluation of existing methods, the decision was made to implement the CART method (Reiter, 2005b)
- CART is a method to model a dependent variable conditionally to a set of predictor variables.
- We have fully synthesized 3 versions of the data for our three GSDS
- Final product will contain 30 synthesis datasets for each GSDS
- We are currently evaluating the research utility and disclosure risk of the three versions of the 3 synthetic data sets


1. Gold Creation
1.1 Data Study
1.2 Evaluation of Existing Research Questions
1.3. End User input about research questions and methods
1.4. Definition of Cohort and Variables
1.5 Pass stakeholder review?
No
Yes
GSDS
2. Synthesization
SDS
3. Evaluation
3.1. GSDS Utility Assessment
3.2 Synthetic Data Research Utility Assessment
3.3 Disclosure Risk Assessment
3.4 Governing Board Approval

Evaluation of synthetic data

- Synthetic data research utility assessment
    - *Do you get the "right" answer from the synthetic data?*

- Disclosure risk assessment
    - *Do the synthetic data pose a risk of disclosure?*

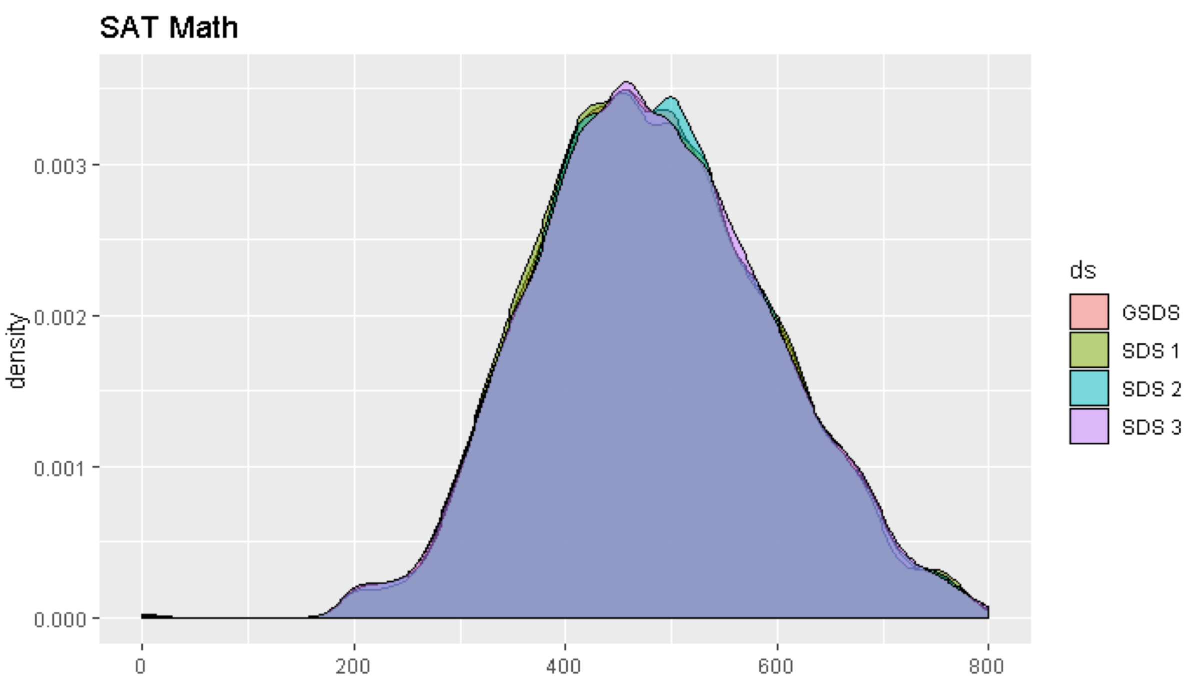
Scope of GSDS

- GSDS is comprised of data from:
  - High school students that entered the workforce
  - High school students that enrolled in post-secondary programs
  - Post-secondary students that entered the workforce

- In total, ~ 100 unique variables in the GSDS
  - Measures for many aspects of education in high school and post-secondary programs
  - Repeated measures for individuals on many variables over time (e.g., GPA, wages)

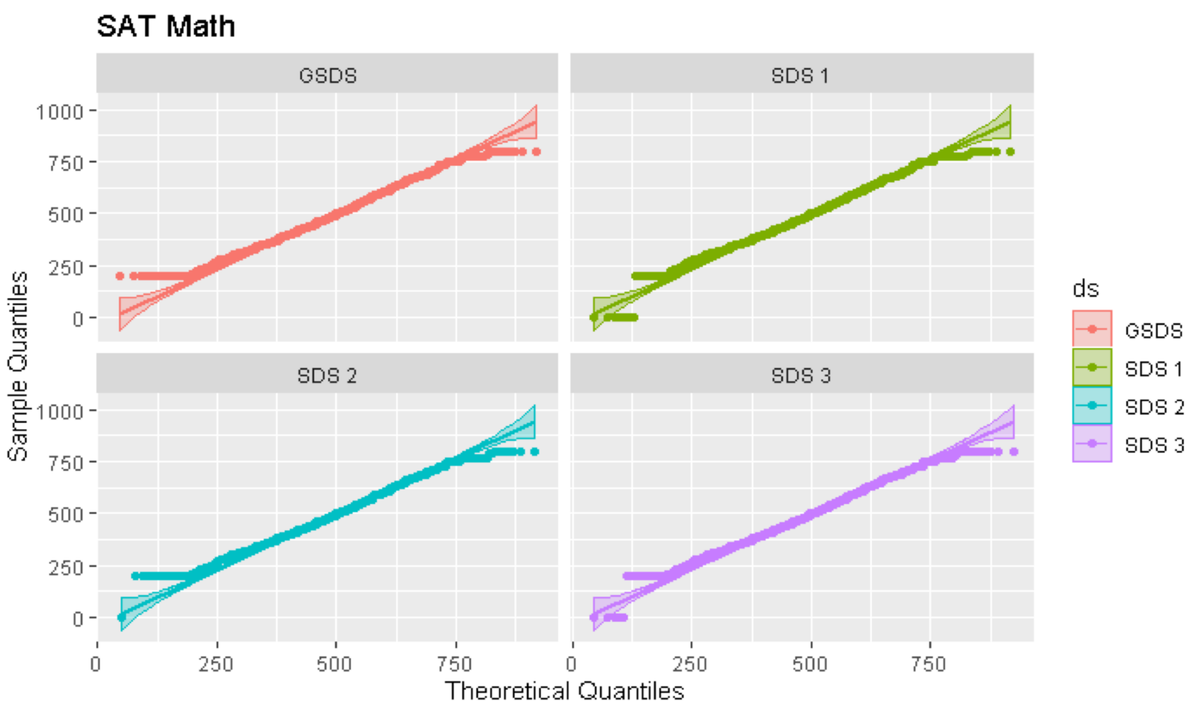
Utility Assessment

- Comparisons of variable distributions
  - Histograms and density plots



Utility Assessment

- Comparisons of variable distributions
  - Quantile plots



Utility Assessment

- Comparisons of descriptive statistics
  - Means and standard deviations
  - Ranges for continuous, factor levels for categorical
  - Proportions of missing values
  - Correlations, contingency tables

- Evaluate within subgroups (e.g., Male/Female)

Utility Assessment - Specific

- How well does synthetic data reproduce the results of specific analyses?

- Gold standard analyses
  - Standardized mean differences
  - Bivariate correlations
  - Multiple regression
  - Logistic regression
  - Time series

Utility Assessment - Specific

- To illustrate components of specific utility assessment, we use a subset of the PS->WF GSDS and three SDSs.
- Regressed (log transformed) 2016 wages on gender, SAT-Math, transformed 2015 wages, and race/ethnicity categories
- The sample size of this cohort was 51,863 students
- We calculate the standardized difference between the estimates of interest based on the GSDS and for each SDS as

$$SD = \frac{\beta_{SDS} - \beta_{GSDS}}{SE_{GSDS}}$$

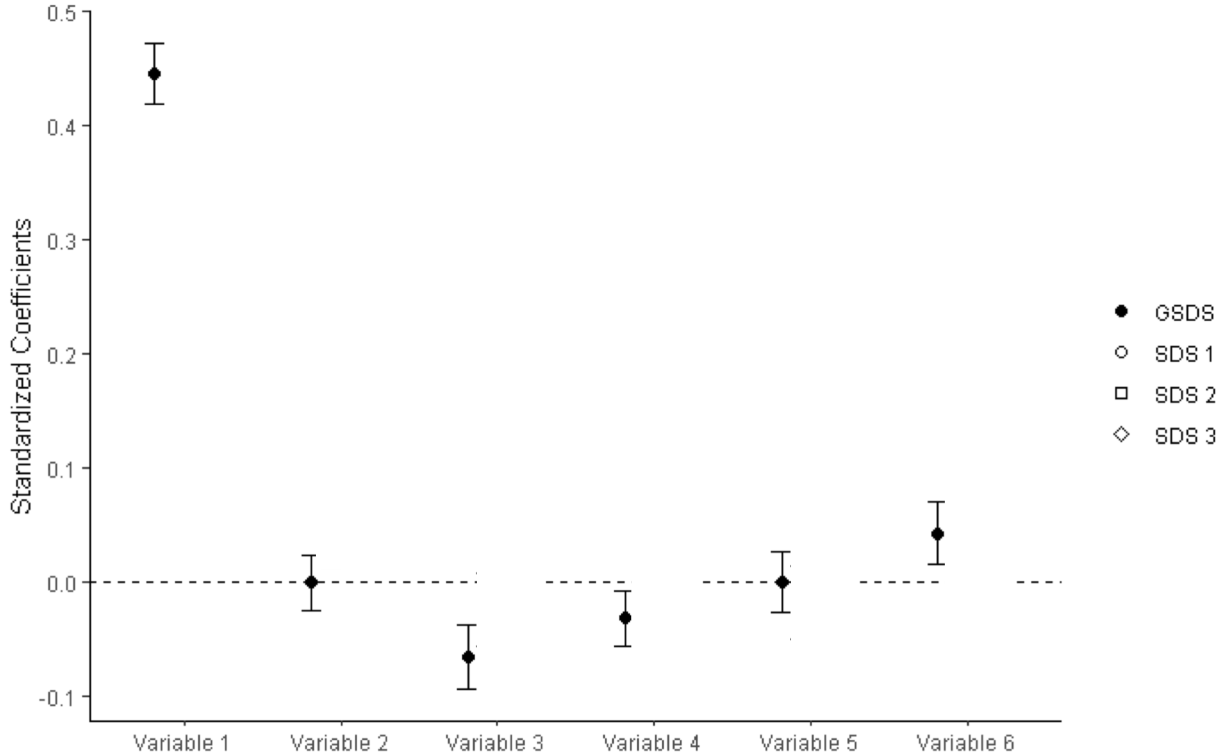
Utility Assessment - Specific

- We also calculate the measure of confidence interval overlap for each estimate (Karr, Kohnen, Organian, Reiter, & Sanil, 2006) as

$$IO = .5\left\{\frac{\min(UCL_{SDS}, UCL_{GSDS}) - \max(LCL_{SDS}, LCL_{GSDS})}{UCL_{GSDS} - LCL_{GSDS}} + \frac{\min(UCL_{SDS}, UCL_{GSDS}) - \max(LCL_{SDS}, LCL_{GSDS})}{UCL_{SDS} - LCL_{SDS}}\right\}$$

- where UCL_{SDS} and LCL_{SDS} represent, respectively, the average upper and lower confidence limits for the replicated estimates based on the SDSs and where UCL_{GSDS} and LCL_{GSDS} are the confidence limits for the estimate based on the GSDS
- Note that when the two confidence intervals do not overlap, the further they are away from each other the more negative the *IO* estimate will become.

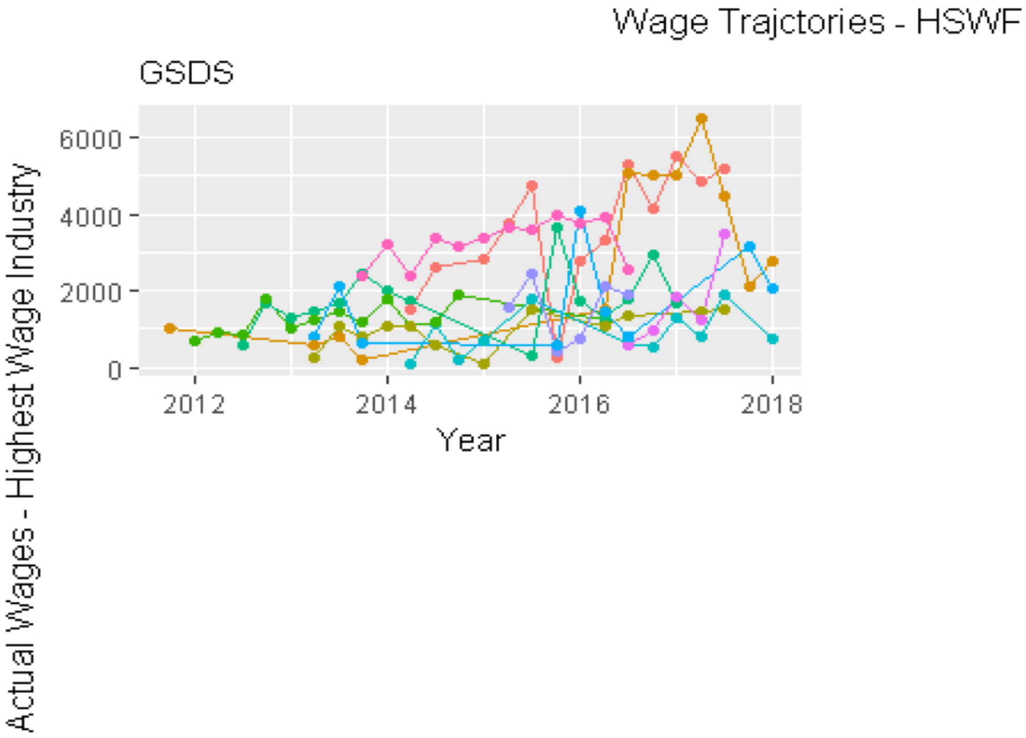
Utility Assessment - Specific



Utility Assessment - Specific

Predictors	GSDS *B* (SE)	AVG SDS *B* (SE)	SD	IO
Variable 1	0.446 (0.014)	0.343 (0.033)	7.572	-0.152
Variable 2	0.001 (0.012)	0.047 (0.014)	3.823	0.107
Variable 3	-0.065 (0.014)	-0.001 (0.018)	4.526	-0.018
Variable 4	-0.031 (0.012)	-0.007 (0.015)	1.912	0.568
Variable 5	0.001(0.014)	-0.004 (0.015)	0.358	0.914
Variable 6	0.043 (0.014)	0.01 (0.016)	2.365	0.443

Utility Assessment - Specific



Utility Assessment - Specific

- Cart model was not well tuned for wages
- Only one lag was used for employment in each sector
- Quarterly wage by sector was creating sparse data

- The solution that was implemented is the following:
  - All possible lags for wages are now used in the predictor set
  - Yearly global wage is synthesized first with all lags
  - then quarterly percentages with all lags
  - then sector percentage within quarterly with same sector lags and all quarters

Utility Assessment - Specific

Year

Utility Assessment - General

- How well does the synthetic data reproduce the variable relationships in the GSDS
  - Not tied to a specific analysis

- Several methods have been proposed
  - Kullback-Leibler divergence
  - Cluster analysis
  - Propensity scores

Utility Assessment - General

- Propensity score method

	Dataset	Subj ID	Variable 1	Variable 2	Variable 3
Real Data	0	1	1	9	3
	0	2	0	12	5
	0	3	0	4	1
	0	4	1	6	1
	...	...	...	...	...
Synthetic Data	1	S1	1	10	0
	1	S2	0	12	0
	1	S3	0	5	0
	1	S4	1	4	0

MLDS Center
Maryland Longitudinal Data System

Utility Assessment - General

- Propensity score estimation
- Logistic regression
    - Interaction terms for higher-order moments
    - Generalized additive model
- Nonparametric classifier
    - CART
        - Naturally models interactions
    - Ensemble methods
        - Random forest
        - Boosted trees

Utility Assessment - General

- Overall measure of utility (Snoke, 2018; Woo, 2009)
    - Mean square error of propensity scores (pMSE)
        - pMSE → 0, less discrepancy between real and synthetic datasets
    - Mostly used for comparing data synthesis methods
- Variable importance
    - Variables with high importance indicate discrepancies between the GSDS and SDS

Disclosure Risk Assessment

- Identification disclosure
  - *relates to the potential for an intruder to match a given record with a specific individual*

- Attribute disclosure
  - *refers to the possibility that even aggregate data collected from these systems have the potential to disclose aspects of different subpopulations that may be sensitive in nature*

Assessing Risk: Identification Disclosure

- Identification Disclosure rests on the assumption that the synthesized data contains identifiable information about individuals from the GSDS on which it was modeled

- For fully synthesized data the “cases” do not exist (there are no “real” records), so theoretically, there is no risk of identity disclosure (the probability would conservatively be 1/N)

- One way to examine identification disclosure in fully synthesized data is to see if it is possible to determine if a specific record from the GSDS is in the SD

High School Cohort

Category	Disclosure Risk
Overall Disclosure Risk	0.000002
Disclosure Risk for Average Person (records near the median across categories)	0.000029
Known NAIC codes (NAIC=22 Utilities Sector)	0.001314
Population Uniques - there are 824 instances where the array of characteristics are unique - 284 of those have no counterpart in the synthetic data at all - 651 have no counterpart in at least one of the synthetic runs	0.209951

Assessing Risk: Attribute Disclosure

- Attribute Disclosure relies on utilizing outside information (such as an additional dataset) to create inferences as a means to identify at-risk groups (<10)
- To assess the attribute disclosure risk we are using a subset of the original GSDS as our "outside source" of information
- The use of the original data provides a worst case scenario of external information an intruder might possess
- Disclosure risk is calculated as the odds of determining sensitive information (such as wages or test scores) using a process of probability matching between the synthetic and "outside" data

Disclosure Risk Assessment

- The below table examines the probability of identification of specific records in the synthesized data given specific levels of knowledge by an intruder. The information in the table is for demonstration purposes.
- The probabilities in the table were developed based on the methodology that is being utilized to calculate the disclosure risk for the synthetic data project but is based on simulations using 51,106 individual records from the Current Population Survey as described in a manuscript by Jerome P. Reiter (2005).
- The probabilities are calculated by dividing 1 over the total number of records identified as having the known characteristics.

Disclosure Risk Assessment

Probabilities of Identification of a specific Record in Synthesized Data[1]	Intruder knows…			
	Demographic Characteristics	Demographic Characteristics and Educ outcomes	Demographic Characteristics, Educ. Outcomes, and Wages	The individual is unique within the source data.
Intruder knows a specific record of interest is in the dataset[2].	0.00045	0.00069	0.00097	0.0047
Intruder does not know a specific record of interest is in the dataset[2] and has knowledge of the underlying process used to synthesize data.	0.0016	0.0028	0.0088	0.01

Summary

- Public release of synthetic data has the potential to both create a safe and robust strategy to comply with data release requests, and, substantially expand access to the MLDS
- Research Utility
  - Multiple methods of assessment
  - Results inform data synthesis model
- Disclosure Risk
  - Identification and attribute disclosure
  - Because all variables are synthesized, in general disclosure risk is low

Next Steps

- Request permission from MLDSC Governing Board for Beta Testing for Research Utility and Disclosure Risk
- Incorporate lessons learned form RU and DR assessment and Beta testing into models
- Seek permission to release synthetic data from MLDSC GB to release synthetic data
- If given permission, build infrastructure and portal to create and maintain synthetic data and disseminate those data

Thank you!

- Contributors:
  - Daniel Bonnery, Yi Feng, Angie Henneberger, Tessa Johnson, Mark Lachowicz, Bess Rose, Terry Shaw, Laura Stapleton, Mike Woolley
  - Email: mlachowi@umd.edu
    mwoolley@ssw.umaryland.edu

- Acknowledgement:
  - This presentation was prepared by the Research Branch of the Maryland Longitudinal Data System Center (MLDSC) as part of funding from the U.S. Department of Education (R372A150045)
  - The Research Branch would like to thank the entire staff of the MLDSC for their assistance with the work and the presentation.

MLDS Center
Maryland Longitudinal Data System

PLEASE COMPLETE THE EVALUATION!!

- 4 ways:
  - Paper forms (submission boxes near registration table)
  - Go to https://www.surveymonkey.com/r/2019STATS-DC
  - Conference app under “Evaluation Forms”
  - Scan the QR code:

Session Number: 12-E